

# Feature Extraction and Analysis of Breast Cancer Specimen

Debnath Bhattacharyya<sup>1</sup>, Tai-hoon Kim<sup>2</sup>, Samir Kumar Bandyopadhyay<sup>3</sup>

<sup>1</sup>Computer Science and Engineering Department,  
Heritage institute of Technology, Kolkata-700107, India

<sup>2</sup>Hannam University, Daejeon – 306791, Korea

<sup>3</sup>Department of Computer Science and Engineering,  
University of Calcutta, Kolkata-700009, India

**Abstract-** In this paper, we propose a method to identify abnormal growth of cells in breast tissue and suggest further pathological test, if necessary. We compare normal breast tissue with malignant invasive breast tissue by a series of image processing steps. Normal ductal epithelial cells and ductal / lobular invasive carcinogenic cells also consider for comparison here in this paper. In fact, features of cancerous breast tissue (invasive) are extracted and analyses with normal breast tissue. We also suggest the breast cancer recognition technique through image processing and prevention by controlling p53 gene mutation to some greater extent.

## 1. Introduction

Imaging techniques play an important role in helping perform breast biopsies, especially of abnormal areas that cannot be felt but can be seen on a conventional mammogram or with ultrasound. One type of needle biopsy, the stereotactic-guided biopsy, involves the precise location of the abnormal area in three dimensions using conventional mammography. Stereotactic refers to the use of a computer and scanning devices to create three-dimensional images. A needle is then inserted into the breast and a tissue sample is obtained. Additional samples can be obtained by moving the needle within the abnormal area [2].

Another type of needle biopsy uses a different system, known as the Mammotome breast biopsy system. The FDA (Food and Drug Administration) approved Mammotome in 1996; the hand-held version of the Mammotome received FDA clearance in September 1999. A large needle is inserted into the suspicious area using ultrasound or stereotactic guidance. The Mammotome is then used to gently vacuum tissue from the suspicious area. Additional tissue samples can be obtained by rotating the needle. This procedure can be performed with the patient lying on her stomach on a table. If the hand-held device is used, the patient may lie on her back or in a seated position.

There have been no reports of serious complications resulting from the Mammotome breast biopsy system. Women interested in this procedure should talk with their doctor.

Digital mammography is a technique for recording x-ray images in computer code instead of on x-ray film, as with conventional mammography. The images are displayed on a computer monitor and can be enhanced (lightened or darkened) before they are printed on film. Images can also be manipulated; the radiologist can magnify or zoom in on an area. From the patient's perspective, the procedure for a mammogram with a digital system is the same as for conventional mammography [2].

Digital mammography may have some advantages over conventional mammography. The images can be stored and retrieved electronically, which makes long-distance consultations with other mammography specialists easier. Because the images can be adjusted by the radiologist, subtle differences between tissues may be noted. The improved accuracy of digital mammography may reduce the number of follow up

procedures. Despite these benefits, studies have not yet shown that digital mammography is more effective in finding cancer than conventional mammography.

The first digital mammography [1] system received U.S. Food and Drug Administration (FDA) approval in 2000. An example of a digital mammography system is the Senographe 2000D. Women considering digital mammography should talk with their doctor or contact a local FDA-certified mammography center to find out if this technique is available at that location. Only facilities that have been certified to practice conventional mammography and have FDA approval for digital mammography may offer the digital system. Many more techniques are available other than the cytogenetic processes, however, these are imaging technologies to detect, diagnose, and characterize breast.

## 2. Previous works

Numerous promising approaches are coming up, few of those only stated here out of our study, and these are very recent.

V. Mallapragada, et al, October, 2007, presented [3, 7] a new concept for real-time manipulation of a tumor using a robotic force controller that monitored the image of the tumor to generate appropriate force to position the tumor at a desired location. The idea was to demonstrate that it was possible to manipulate a tumor in real-time by applying controlled external force in an automated way such that the tumor did not deviate from the path of the needle. The success of this approach had the potential to reduce the number of attempts a surgeon make to capture the desired tissue specimen, minimized tissue damage, improved speed of biopsy, and reduced patient discomfort.

Cigdem Gunduz, et al, 2004, reported a computational method that modeled a type of brain cancer using topological properties of cells in the tissue image. They constructed the graphs based on the locations of cells within the image. They used the Waxman model in their experiment [4].

C. Cagatay Bilgin, et al, 2007, classified [5] the breast cancer tissues using graph theory. Image segmentation approach was used and Euclidean Distances were calculated between vertices [5]. Cell Graphs were generated by considering the cell locations. Approach was same to the greater extent with the work of Cigdem Gunduz, et al, 2004.

These approaches toward automatic detection of cancer were actually failed because the types of cancers identified more and more.

A.M. Tang, et al, 2008, proposed, simultaneous capturing of ultrasound (US) and magnetic resonance (MR) images allowed fusion of information obtained from both modalities. An MR-compatible US system where MR images were acquired in a known orientation with respect to the US imaging plane and concurrent real-time imaging could be achieved. Compatibility of the two imaging devices was a major issue in the physical setup. Tests were performed to quantify the radio frequency (RF) noise introduced in MR and US images, with the US system used in conjunction with MRI scanner of different field strengths (0.5 T and 3 T). Furthermore, simultaneous imaging was performed on a dual modality breast phantom in the 0.5 T open bore and 3 T close bore MRI systems to aid needle-guided breast biopsy. Fiducial based passive tracking and electromagnetic based active tracking were used in 3 T and 0.5 T, respectively, to establish the location and orientation of the US probe inside the magnet bore. Their results indicated that simultaneous US and MR imaging were feasible with properly-designed shielding, resulting in negligible broadband noise and minimal periodic RF noise in both modalities. US could be used for real time display of the needle trajectory, while MRI could be used to confirm needle placement [6].

C. Zhu, et al, 2009, have explored [8] the use of a fiber-optic probe for in vivo fluorescence spectroscopy of breast tissues during percutaneous image-guided breast biopsy. A total of 121 biopsy samples with accompanying histological diagnosis were obtained clinically and investigated in their study. The tissue spectra were analyzed using partial least-squares analysis and represented using a set of principal

components (PCs) with dramatically reduced data dimension. For nonmalignant tissue samples, a set of PCs that account for the largest amount of variance in the spectra displayed correlation with the percent tissue composition. For all tissue samples, a set of PCs was identified using a Wilcoxon rank-sum test as showing statistically significant differences between: 1) malignant and fibrous/benign; 2) malignant and adipose; and 3) malignant and nonmalignant breast samples. These PCs were used to distinguish malignant from other nonmalignant tissue types using a binary classification scheme based on both linear and nonlinear support vector machine (SVM) and logistic regression (LR). For the sample set investigated in this study, the SVM classifier provided a cross-validated sensitivity and specificity of up to 81% and 87%, respectively, for discrimination between malignant and fibrous/benign samples, and up to 81% and 80%, respectively, for discriminating between malignant and adipose samples. Classification based on LR was used to generate receiver operator curves with an area under the curve (AUC) of 0.87 for discriminating malignant versus fibrous/benign tissues, and an AUC of 0.84 for discriminating malignant from adipose tissue samples. This study demonstrated the feasibility of performing fluorescence spectroscopy during clinical core needle breast biopsy, and the potential of that technique for identifying breast malignancy in vivo.

Lin Yang, et al, 2007, introduced a Grid-enabled CAD to perform automatic analysis of imaged histopathology breast tissue specimens [10]. More than 100,000 digitized samples (1200 × 1200 pixels) were processed on the Grid. They analyzed results for 3744 breast tissue samples, which were originated from four different institutions using diaminobenzidine (DAB) and hematoxylin staining. Both linear and nonlinear dimension reduction techniques were compared, and the best one was applied to reduce the dimensionality of the features. The results shown that the Gentle Boosting using an eight node CART decision tree as the weak learner provided the best result for classification. The algorithm has an accuracy of 86.02% using only 20% of the specimens as the training set.

### 3. Our work

We used free Tissue Blocks downloaded from OriGene Technologies, Inc, 2009 [9]. Here in our experiment, 18 invasive breast cancer tissues from different 18 patients and 8 non-cancerous falsely detected breast tissues from 8 different normal females are considered. Each of the 24-bit BMP Image size is 640 x 480 Pixels.

#### 3.1. 24-bit Color Image to 256-color Gray Image

1. Take this 24-Bit BMP file as Input file and open the file in Binary Mode, (Size  $M \times M$ ).
2. Copy the Image Info (First 54 byte) of the Header from Input 24-Bit Bmp file to a newly created BMP file and edit this Header by changing filesize, Bit Depth, Colors to confirm to 8-Bit BMP.
3. Copy the Color Table from a sample gray scale Image to this newly created BMP at 54th Byte place on words.
4. Convert the RGB value to Gray Value using the following formula:
  - a.  $\text{blueValue} = (0.299 * \text{redValue} + 0.587 * \text{greenValue} + 0.114 * \text{blueValue});$
  - b.  $\text{greenValue} = (0.299 * \text{redValue} + 0.587 * \text{greenValue} + 0.114 * \text{blueValue});$
  - c.  $\text{redValue} = (0.299 * \text{redValue} + 0.587 * \text{greenValue} + 0.114 * \text{blueValue});$
  - d.  $\text{grayValue} = \text{blueValue} = \text{greenValue} = \text{redValue};$
5. Write to new BMP file.

Take 24-bit BMP color image as input. Then convert it to 256-color Gray Scale image by following this algorithm. This 256-color Gray Scale image is the output of the algorithm. In this algorithm, first read the red, blue and green value of each pixel and then after formulation, three different values are converted into gray value, stated in Step 4.

### 3.2. 256-color Gray Image to Bi-color (using Pixel Clustering on Threshold Value, T)

1. Open 256-color Image (Size  $M \times M$ )
2. Read a Pixel value
3. If the Pixel Intensity value less than or equal to T (128) then make it 0 Else make it 255 and write into same Pixel Location
4. Go to Step 2 until end of file
5. Close file

This algorithm is actually used here to convert the Gray Image to Bi-color (Monochrome Image). In some cases we can say this is the Edge Detection Algorithm set on a Threshold Value.

### 3.3. Cell Representation Algorithm on Spatial Domain

1. Open Bi-color Image (Size  $M \times M$ )
2. Set a 2D Integer Array (equivalent to size of Bi-color Image,  $M \times M$ )
3. Read a Pixel value
4. Store corresponding location of 2D Array (If the Pixel value is 255, make it 1 in our case)
5. Go to Step 2 until end of file
6. Close file
7. Draw the Graph on 2D Space using that generated Binary Matrix
8. End

The Generated Binary Matrix can be used for future statistical analysis to make the system automatic, definitely, with other biological characteristics of Breast Cancer Cells. Here in this work we compare the those Graphs and suggest for further pathological test or for no need of test.

## 4. Analysis and Result

Here the challenge is Mammogram and Digital Biopsy. Problem with mammogram may arise biopsy also. Now we are considering some kind of mammogram analysis. We have noticed same problem with Biopsy. In most individuals the bulk of the breast extends from the second to the seventh rib. Since breast tissues often curve around the lateral margin of the pectoralis major muscle (Figure 1), the orientation of the muscle is important for optimal mammographic positioning. The pectoralis major muscle spreads like a fan across the chest wall. Portions of the pectoralis major muscle attach to the clavicle, the lateral margin of the scapula, costal cartilage and the aponeurosis of the external oblique muscles of the abdomen. All these fibers converge on and attach to the greater tubercle of the humerus. The free fibers predominantly run obliquely over the chest from the medial portion of the thorax toward the humerus. The relationship of the breast to the pectoralis major muscle influences two-dimensional projectional imaging, such as mammography. Since the breast tissue is closely applied to the muscle, some of the lateral tissues can only be imaged through the muscle. As with any soft-tissue structure overlying muscle, it is easier to project the breast into the field of view by pulling it away from the chest wall and compressing it with the plane of compression along the obliquely oriented muscle fibers of the pectoralis major muscle. In order to maximize the tissue imaged, the free portion of the muscle should be included in the field of view.

In view of the enormous amount of work that has been done in an effort to understand the breast and the development of breast cancer, it is surprising that the normal breast has never been clearly defined. This is likely due to the fact that since breast cancer is really the only significant abnormality that occurs in the breast, it is really only the changes that appear to predispose to breast cancer that are considered significant. There is a large range of histologic findings that occur in women who never develop breast

cancer, but where normal ends and abnormal begins is not obvious, and past classifications have been found to be inaccurate.

The ability to detect breast cancers earlier requires high-quality imaging, proper film processing, systematic review of the images, reasoned interpretation, the ability to solve problems raised by the imaging, and the ability to guide the diagnostic removal of cells or tissue for diagnosis. The interpreter should participate in all aspects of this process. It is very important that quality control be supervised by the interpreter(s) of the images so that any image degradation can be detected and corrected as quickly as possible.

Errors can be reduced by following a carefully structured approach to the process. The detection and diagnosis of breast cancer can be divided into five very specific tasks: Detection—Find it. Verification—Is it real? Triangulation—Where is it? Identification—What is it? Management—What should be done about it?



Figure 1. Computed Breast tomography with the breasts in the pendent position shows breast tissue on the left adjacent to the pectorals major muscle extending up toward the axilla.

Ductal Cancer can spread up and down the duct network and remain in situ, whereas invasive cancer can be found associated with a part of the process. This finding would support the continuum theory. Their data suggest that one of the already genetically unstable cells in the duct developed an invasive clone and that this clone proliferated while the remaining in situ cells, unable to invade, continued to proliferate and spread up and down the ducts. This observation explains invasive breast cancer can be found in the same lesion (Figure 2). In figure 2, outside the ducts and lobules a huge amount of breast muscle and tissue are present and here is the challenge.

An understanding of breast tissue patterns as they apply to the sensitivity of mammographic detection of breast malignancy is important. The greater the amount of fat within the breast, the easier it is to recognize a water-density tumor (Figure 3). As in any other x-ray study, the margins of a water-density cancer will be obscured or invisible when they are contiguous with normal tissue of equivalent x-ray attenuation. In breasts in which the parenchyma is nonuniform, the x-ray attenuation will vary in a nonuniform way, making it difficult to detect a small cancer whose margins are similarly nonuniform. In the breast that is heterogeneously dense or extremely dense, the sensitivity of mammography, not only for the early detection of malignancy, but also for large cancers is somewhat diminished because of the difficulty of finding ill-defined cancers within the inhomogeneous background.

The fact that mammography can detect very small cancers but can also miss some very large cancers is confusing to clinicians and the public. Figure 4-11 are useful for explaining how mammography can detect many very small cancers, but some large palpable cancers can still be difficult to image.

The dense breast is not the only reason for overlooking cancers. It is of some interest that among cancers overlooked in the screening study many cancers were overlooked in women with predominantly fatty

breast tissue. Detecting small cancers in the dense breast is more difficult, but early-stage breast cancer can be detected by mammography among these women. In a review of 118 women with breast cancer detected by mammography alone<sup>32</sup>, among women under the age of 50 years, we found that 70% were detected in women with radiographically dense breast tissue and these were at a smaller size and earlier stage than among women with palpable cancers. Even though a higher proportion of younger women have dense tissues, recent data from modern mammography screening programs show that mammography can detect early cancers among women aged 40 to 49 years at the same proportion as for women aged 50 to 59 years<sup>33, 34</sup>. The dense breast does reduce the sensitivity of mammography somewhat, but should not deter screening among these women and is not the sole cause for overlooking breast cancers.

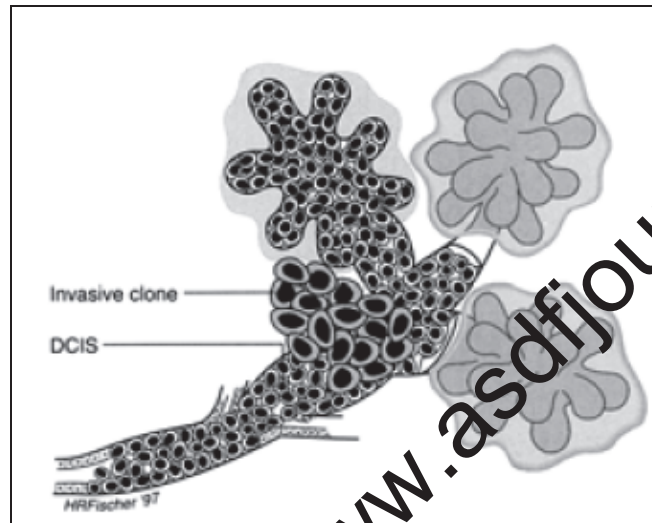


Figure 2. Cells that are proliferating out of control but lack the ability to invade may continue to grow within the duct while a clone that has developed invasive capability can be growing simultaneously in the same lesion.

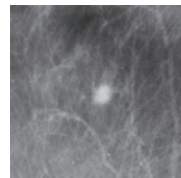


Figure 3. This 6-mm invasive ductal carcinoma is easily visible because it is surrounded by fat tissue.

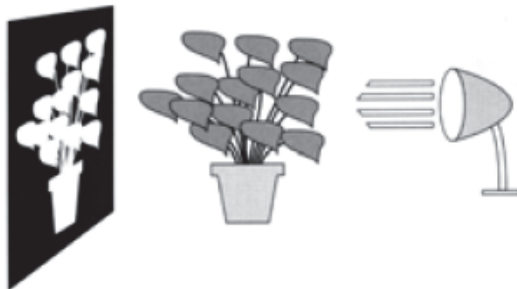


Figure 4.

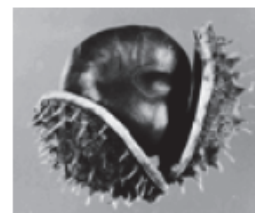


Figure 5

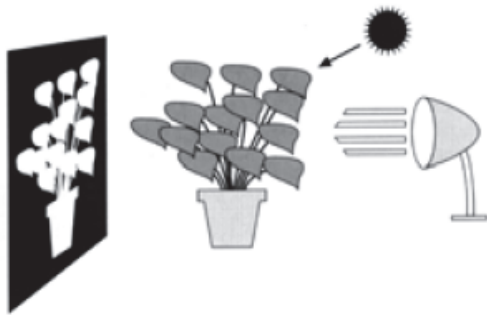


Figure 6.

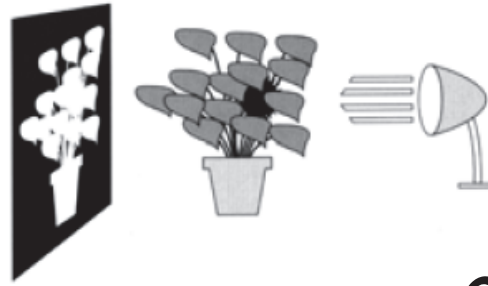


Figure 7.

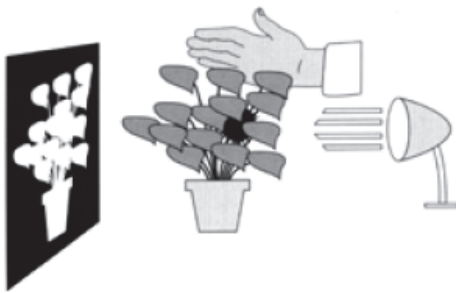


Figure 8.



Figure 9.



Figure 10.

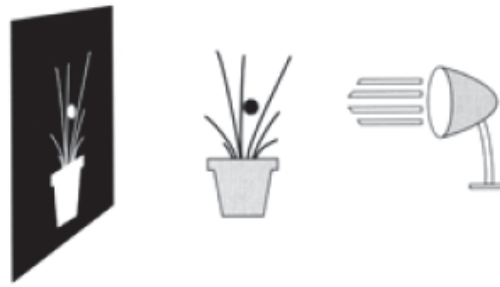


Figure 11.

The projection of a potted plant onto the wall using a spotlight (Figure 4) is a good analogy to the breast and cancer detection by clinical breast examination and mammography. Assume that a chestnut, with its hard shell (Figure 5), is placed in among the branches and leaves of the plant (Figure 6). If the leaves are densely packed, the nut even a very large one may not be visible (Figure 7), yet fingers pressed against it can easily feel it (Figure 8). If the plant has fewer leaves, analogous to the breast with less fibrous tissue, then the nut becomes more visible (Figure 9). If there are few leaves, then even a very small nut is visible (Figure 10), and if an extremely small nut is nestled between the rigid stems of the plant, the nut may be easily visible, but not palpable because it is protected by the stems (Figure 11).

Our algorithms, specially first and second, are used to remove the huge amount of tissue and fat from the Cancerous cells within the biopsy samples, here we are naming these as tissue blocks, shown in Figure 12.

Our target is to get the image something like in Figure 11. Outputs of those algorithms are shown in Figure 12a-c, for normal breast tissue. Figure 12c, shows the cells with black spotted on the space.

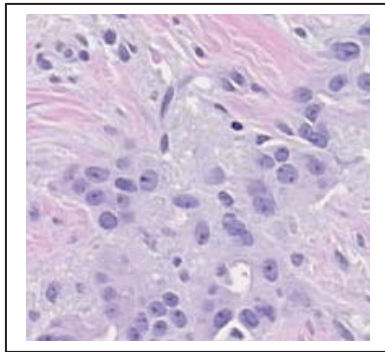


Figure 12a. 24-bit Color Image

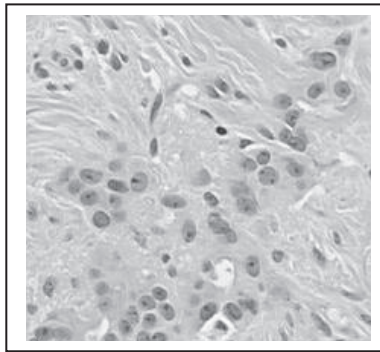


Figure 12b. 256-Color Gray Image



Figure 12c. Bi-color Color Image

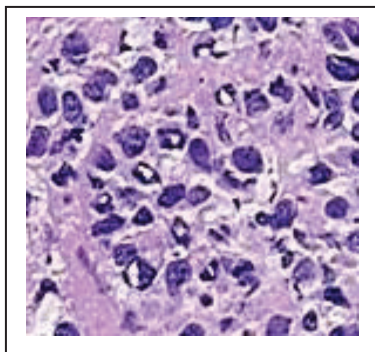


Figure 13a. 24-bit Color Image

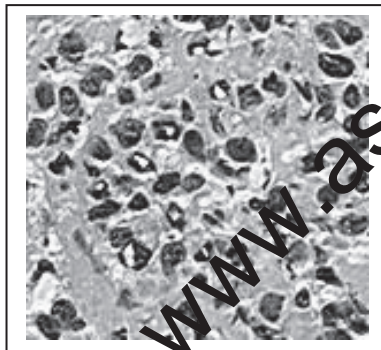


Figure 13b. 256-Color Gray Image

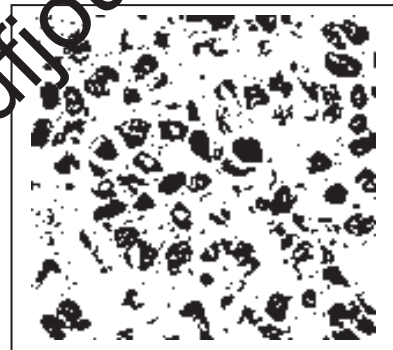


Figure 13c. Bi-color Color Image

Figure 13a-c, shows the Cancerous cells, some kind of abnormal size and numbers are marked. These outputs also from the same set of algorithms. We have conducted the observations using 18 different patients tissue block all are invasive situ breast cancer and 8 normal breast tissue blocks.

Graphical observations also conducted as shown in Figure 14 and 15. In case of normal tissue, disconnected cell graphs have been identified with few numbers. On the other hand, in case of invasive breast cancer tissue uncounted connected cell graphs are observed.

## 5. Conclusion

Till date, it is observed genetic mutation of certain oncogenes is responsible for any type of cancers. Modern techniques are using for treatment and chemotherapy is an established way of controlling cancers now a days. But, the question is that "Why the oncogenes are suddenly changing their behavior or becoming inactivated"? Next we will put more effort on genetical behavior of cancer genes and how these can be tuned that leads to more biometric.

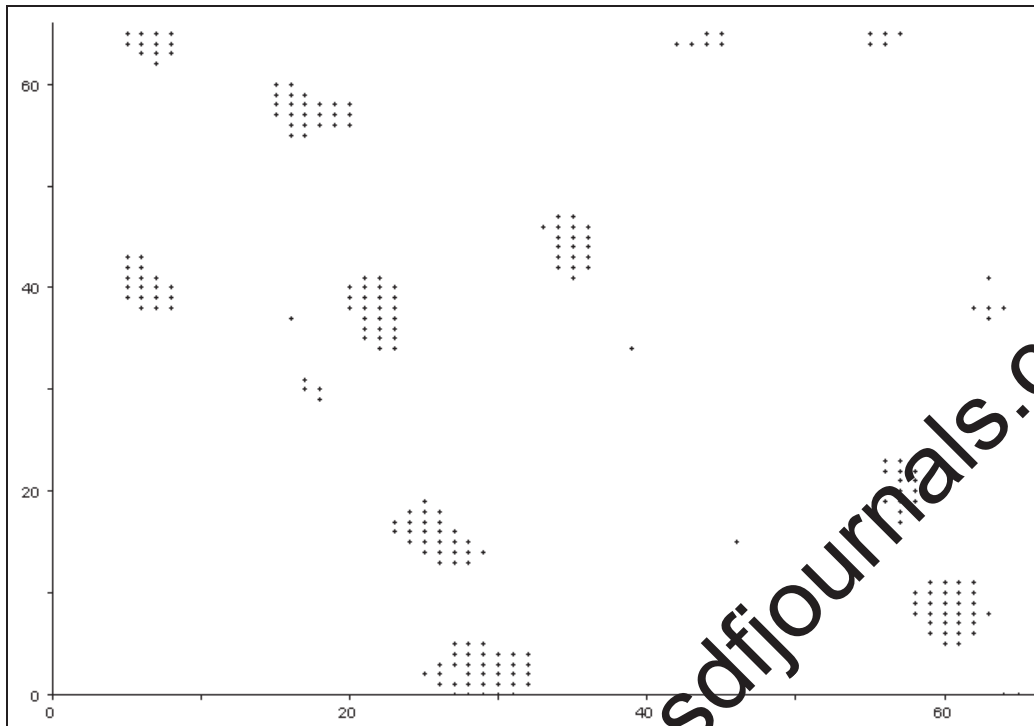


Figure 14. Normal breast tissue with normal cells in graphical problem space with dotted signs.

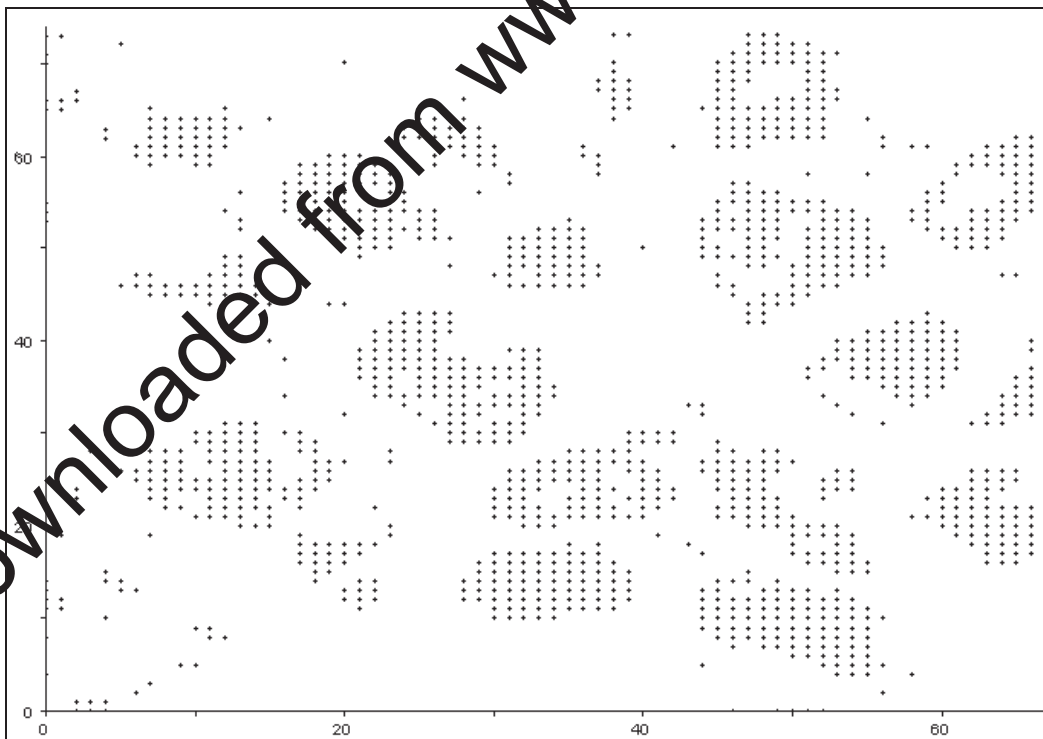


Figure 14. Invasive situ breast cancer tissue with cells in graphical problem space with dotted signs.

## 6. Acknowledgement

This work was supported by the Security Engineering Research Center, granted by the Korea Ministry of Knowledge Economy. And this work has successfully completed by the active support of Prof. Tai-hoon Kim, Hannam University, Republic of Korea.

## References

- [1] FDA Web site, <http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfMQSA/mqsa.cfm> [last visited on September 30, 2009].
- [2] National Cancer Institute (NCI) Web site, <http://www.cancernet.gov> [last visited on September 30, 2009].
- [3] V. Mallapragada, N. Sarkar, and T.K. Podder, "A Robotic System for Real-time Tumor Manipulation During Image guided Breast Biopsy", IEEE International Conference on Bioinformatics and Bioengineering, October 14-17, 2007, Boston, MA, pp. 204-210.
- [4] Cigdem Gunduz, Bulent Yener, and S. Humayun Gultekin, "The cell graphs of cancer", Bioinformatics, Oxford University Press, Vol 20, Issue 1, January, 2004, pp. 145-151.
- [5] C. Cagatay Bilgin, Cigdem Demir, Chandandeep Nagi, and Bulent Yener, "Cell-Graph Mining for Breast Tissue Modelling and Classification", 29th IEEE EMBS Annual International Conference, August 23-26, 2007, Lyon, France.
- [6] A.M. Tang, D.F. Kacher, E.Y. Lam, K.K.Wong, F.A. Jolez, and E.S. Yang, "Simultaneous Ultrasound and MRI System for Breast Biopsy: Compatibility Assessment and Demonstration in a Dual Modality Phantom", IEEE Transactions on Medical Imaging, Vol. 27, Issue 2, February 2008, Davis, CA, USA, pp. 247-254.
- [7] V. Mallapragada, N. Sarkar, and T.K. Podder, "Robot-Assisted Real-Time Tumor Manipulation for Breast Biopsy", IEEE Transactions on Robotics, Vol. 25, Issue 2, April 2009, pp. 316-324.
- [8] C. Zhu, E.S. Burnside, G.A. Sisney, R. Salkowski, J.M. Harter, B. Yu, and N. Ramanujam, "Fluorescence Spectroscopy: An Adjunct Diagnostic Tool to Image-Guided Core Needle Biopsy of the Breast", IEEE Transactions on Biomedical Engineering, Vol. 56, Issue 10, October 2009, pp. 2518 - 2528.
- [9] Origene Web site, <http://www.origene.com/> (one such example we have used here, <http://www.origene.com/assets/images/tissues/blocks/CU0000005705.AF1.20X.jpg>) [last visited on August 29, 2009].
- [10] Lin Yang, Wenjin Chen, Peter Meer, Gratian Salaru, Michael D. Feldman, and David J. Foran, "High Throughput Analysis of Breast Cancer Specimens on the Grid", MICCAI 2007, Part I, October 29 - November 4, 2007, Brisbane, Australia, LNCS 4791/2007, pp. 617-625.